Report on IHMC- CMU-Pitt Research

Executive Summary

NRA A2-37143

"Automated Discovery Procedures for Gene Expression and Regulation from Microarray and Serial Analysis of Gene Expression Data"

NCC 2-1295

"Multi-Domain Network Learning Algorithms of Latent Variable Interpretation and Discovering Genetic Regulation"

April 2001 – April 2002

http://www.phil.cmu.edu/projects/genegroup

Research Team

- William Buckles (Ph.D, Professor, Tulane)
- Tianjiao Chu (Ph.D Student, Logic, Methodology and Computation, CMU)
- Greg Cooper (M.D. Ph.D Associate Professor, School of Medicine, Pitt
- David Danks (Ph.D, Research Scientist, IHMC)
- Clark Glymour (Ph.D, P.I., Senior Resarch Scientist and John Pace Scholar, IHMC; Alumni University Professor, CMU)
- Dan Handley (M.S. Student, Logic, Methodology and Computation, CMU
- Subramani Mani (Ph.D Student, Biomedical Informatics, Pitt)
- Rob O'Doherty (Ph.D ,Assistant Professor, School of Medicine, Pitt)
- Dave Peters (Ph.D , Human Genetics, Pitt
- Joseph Ramsey (Ph.D, Research Programmer, CMU)
- Jamie Robins (M.D. School of Public Health, Harvard)
- Raul Saavedra (Ph.D, Student, Computer Science, Tulane)
- Richard Scheines (Ph.D, Associate Professor, CMU)
- Nicoleta Servan (Ph.D Student, Statistics, CMU)
- Ricardo Silva (Ph.D student, Computer Science, CMU)
- Peter Spirtes (Ph.D, Research Scientist IHMC; Professor, CMU)
- Larry Wasserman (Ph.D, Professor, CMU)
- Frank Wimberly (Ph.D, Research Programmer, IHMC)
- Changwon Yoo (Ph.D Student, Biomedical Informatics, Pitt)

Two Related Goals

- Investigating the prospects for more rapid and accurate determination of genetic regulatory networks using recently developed technologies (microarrays and SAGE)
- Investigating the prospects for determining the underlying components of measured phenomena, and the influences such components have on one another

Background on Genetics

- Proteins do most of the work in the cell
- Cell reproduction, metabolism, and responses to the environment are all controlled by proteins
- Each gene is a machine for constructing (approximately) a single protein
- The rate at which a gene constructs proteins is influenced by concentrations of regulator proteins

Gene Regulatory Networks

- Some genes manufacture proteins which control the rate at which other genes manufacture proteins (either promoting or suppressing)
- Hence some genes indirectly (via the proteins they create) regulate other genes, which in turn regulate the operation of the cell
- The system by which genes regulate each other is called the genetic regulatory network, and can be represented by a directed graph (which is a special case of a Bayes network)

Measuring Gene Expression Levels

- A gene's "expression level" is an approximate measure of the concentration of mRNA transcripts and an more indirect measure of the rate of synthesis of corresponding proteins.
- Recently developed technologies--microarrays and Serial Analysis of Gene Expression, or SAGE--allow thousands of gene expression levels to be measured simultaneously
  - The kinds of measurement errors that these technologies introduce is not well understood
  - The best way to use these tools to discover gene regulatory networks is not known

Relevance to NASA

- Gene expression in microgravity has been shown to differ significantly from expression in Earth gravity
  - Understanding gene regulation in plants, animals and humans is likely to be important for long term extraterrestrial habitation
  - Determining regulatory structure is a present laborious, slow and costly
  - Need for systematic study of the reliability and accuracy of scores of proposals for applying statistical/machine learning procedures to speed up the process

Background on Latent Structure Analysis

- Measurements are often of effects of other scientifically interesting variables not directly mesured.
- Number and identity of underlying causal or compositional variables may not be entirely known.
- Measured effects can influence other measured effects (e.g., through between channel signal leakage in multi-channel

Background on Latent Structure Analysis

- With no prior cluster information and with the possibility of measured-measured and latent-latent influences, none of the standard data analysis procedures (e.g., factor analysis, principal components, independent components) give reliable (i.e., asymptotically correct) information about all of:
- Number of latent variables
  - Clustering of measured
  - Causal or compositional relations among latent variables

Relevance to NASA

- NASA collects vast quantities of observational data on the Earth, the solar system and the cosmos, much of it spectral
  - Need for automated, fast, reliable procedures extracting relevant causal information from diverse datasets — procedures that integrate expert knowledge
  - Inadequacy of current methods (model specific, clustering algorithms) for this task
  - Principled procedures using Bayes network methods offer promising alternatives
    - They have succeeded in other spectral applications
    - (J. Ramsey, et al., "Automated Identification of Carbonate Composition from Reflectance Spectra," Data Mining and Knowledge Discovery, in press.)

Structure of the Projects

- Statistical Foundations
  - Multiple testing problem
  - Measurement error models
- Search Algorithms
  - Different kinds of inputs
  - Different assumptions about background knowledge
- Experiments
  - Microarray
  - SAGE
- Testing
  - Application to known genetic regulatory networks
  - Application to simulated data

First Year Results: Algorithms

- Many algorithms for inferring causal networks that have been applied to inferring gene regulatory networks assume the input is associations between measured features of *individuals*
- But microarrays and SAGE measure *average* gene expression levels over many cells rather than for a single cell
- What is the feasibility of inferring regulatory networks from associations between averages?

  - *Feasibility* for linear and local-linear regulatory functions
  - *Impossibility* for the mathematical form of the regulatory function of sea urchin Endo 16 gene, one of the best established.
    - T. Chu, C. Glymour, R. Scheines and P. Spirtes, "A Statistical Problem for Inference to Regulatory Structure form Associations of Gene Expression Measurements with Microarrays" *Bioinformatics*, submitted.

First Year Results: Statistics

- Current methods for determining from SAGE measurements which genes are changing in response to experimental manipulations are incorrect
- Correct method requires estimating additional experimental parameters, and leads to the conclusion that many fewer genes are changing than had been previously thought
- T. Chu, “Computation of Variance in SAGE Measurements of Gene Expression” Technical Report, Logic, Methodology and Computation, 2002.
- Future plan – apply the new method to SAGE measurements of the response of genes to shear stress (data already gathered)

First Year Results: Statistics

- Standard techniques for testing whether a gene expression level has changed due to an experimental manipulation were not designed to be applied to test thousands of genes simultaneously
- Recent developments (False Discovery Rate tests) do allow simultaneous testing of thousands of genes
- Further improvements of the False Discovery Rate procedure have been made
  - C. Genovese, and L. Wasserman, “Bayesian and Frequentist Multiple Testing”, CMU Department of Statistics Technical Report 764, April, 2002.

First Year Results: Algorithms

- Implementation and testing (on simulated data) of a correct (under explicit assumptions) algorithm for causal clustering and for determining latent structure
- R. Silva, CMU Master's Thesis, Center for Automated Learning and Discovery
- Extension to time series of learning algorithms for dynamical Bayes Nets
  - D. Danks, "Constraint-Based Learning Algorithm for Dynamical Bayes Nets, Conference on Uncertainty in Artificial Intelligence," submitted.
- Development and proof of correctness for an improved algorithm for inferring Bayes networks across distinct data sets with overlapping variable sets
  - D. Danks, "Efficient Learning of Bayes Nets from Databases with Overlapping Variables," IHMC Technical Report, 2002.

First Year Results: Algorithms

- Development and testing of algorithms for maximizing information obtained from “knockout” experiments
  - R. Silva, C. Glymour, D. Danks, “Inferring Genetic Regulatory Structure from First and Second Moments,” Technical Report, Logic, Methodology and Computation, 2002.
  - Development, implementation and testing of a genetic algorithm for linear Bayes networks (structural equation models)
  - S. Harwood and R. Scheines, “Learning Linear Causal Structure Equation Models with Genetic Algorithms” (2001) Tech Report CMU-PHIL-128, submitted to Conference on Knowledge Discovery and Data Mining.
  - S. Harwood and R. Scheines, “Genetic Algorithm Search over Causal Models” (2001) Tech Report CMU-PHIL-131, submitted to Conference on Uncertainty in Artificial Intelligence.
  - Development of an algorithm for regulatory structure from mixed observational and knockout data

First Year Results: Testing

- Very few genetic regulatory networks are known, and even fewer details about the functional relationships among the genes are known
- How can the accuracy of a causal discovery algorithm be tested?
- Generate simulated data from made up gene regulatory networks, so that the generating mechanism is known

First Year Results: Testing

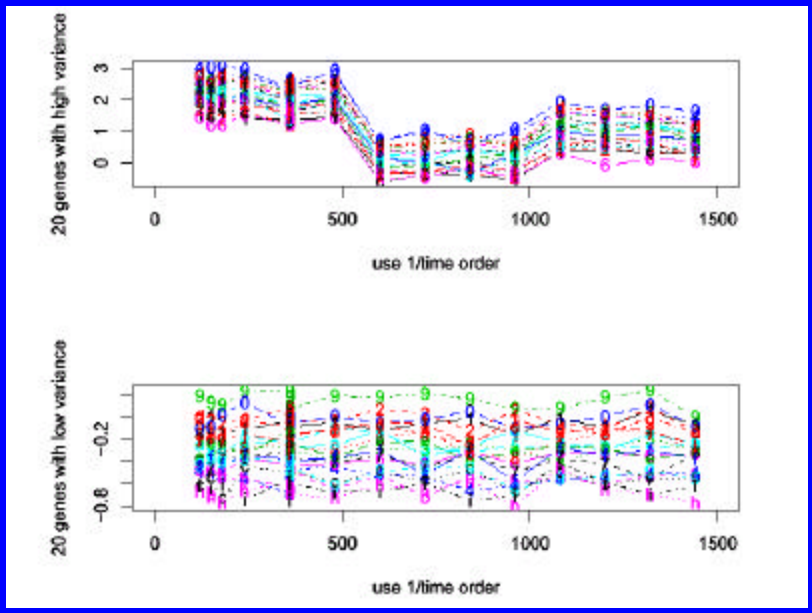
- Implementation of a flexible program for generating simulated microarray data that allows the user to conveniently specify many different
  - Functional relationships between cells
  - Measurement errors
  - Averaging over different numbers of cells
  - Gene regulatory network structures (including varying time lags)
    - J. Ramsey and R. Scheines, (2001) “Simulating Genetic Regulatory Networks,” Technical Report CMU-PHIL-124.
- Implementation of half a dozen algorithms proposed in the literature for inferring regulatory structure from expression associations in microarray measurements (more to be implemented)

First Year Results: Experiments

- Fat cells from mice are treated with troglitazone, which increases the efficiency of the biological actions of insulin in diabetes and obesity
- Which genes are activated?
- Microarray chips used to make 47 measurements of gene expression level at 35 time points for 5355 genes

First Year Results: Experiments

- Normalize data to remove chip-to-chip effects
- Perform statistical tests to determine which genes are changing, adjusting for multiple tests

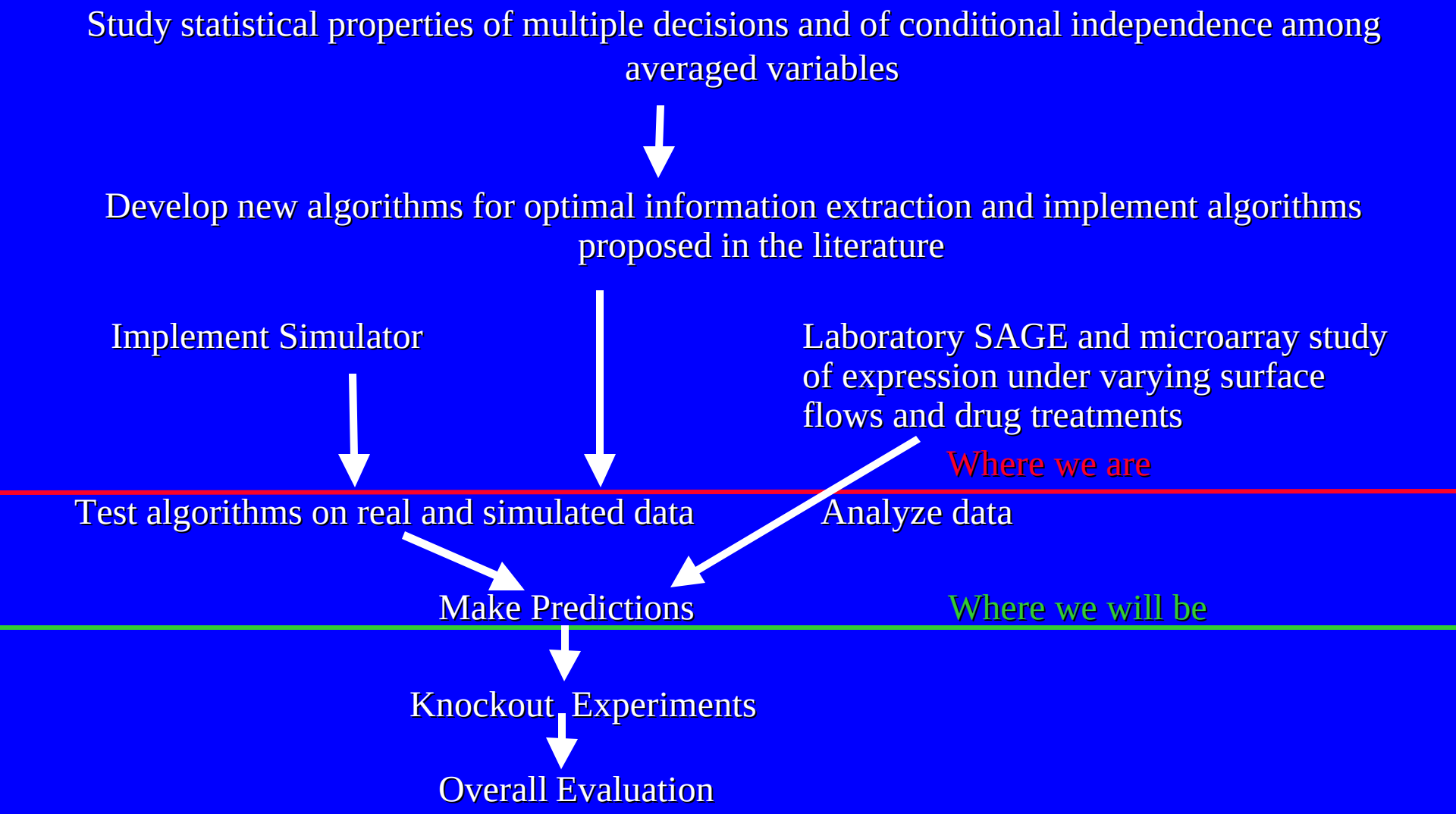


Comparing 20 genes that change most with 20 that change least

Current Work: Experiments

- Remove outlying genes
- Improve the test performed for whether a gene is changing over time
- Introduce clustering methods for data
- Use slower but more accurate measurement techniques (Northern Blots) to
  - Test the hypotheses about which genes change according to the microarray analysis
  - Learn about errors in measurement when using microarrays

Gene Research Plans: May 2002 – May 2003



Latent Structure Research Plans, 2002-2003

- Improve efficiency
- Test on large simulated data sets
- Prove asymptotic correctness
- Investigate non-linear generalizations